

What is claimed is:

- 1 1. A method comprising the steps of:
 - 2 (A) receiving an email message from a simple mail transfer protocol (SMTP) server,
 - 3 the email message comprising:
 - 4 (A1) a 32-bit string indicative of the length of the email message;
 - 5 (A2) a text body;
 - 6 (A3) an SMTP email address;
 - 7 (A4) a domain name corresponding to the SMTP email address;
 - 8 (A5) an attachment;
 - 9 (B) tokenizing the text body to generate tokens representative of words in the text;
 - 10 (C) tokenizing the SMTP email address to generate a token representative of the
 - 11 SMTP email address;
 - 12 (D) tokenizing the domain name to generate a token that is representative domain
 - 13 name;
 - 14 (E) tokenizing the attachment to generate a token that is representative of the
 - 15 attachment, the tokenizing step comprising the steps of:
 - 16 (E1) generating a 128-bit MD5 hash of the attachment;
 - 17 (E2) appending the 32-bit string to the generated MD5 hash to produce a 160-
 - 18 bit number; and
 - 19 (E3) UUencoding the 160-bit number to generate the token representative of the
 - 20 attachment;
 - 21 (F) determining a probability value for each of the generated tokens;
 - 22 (G) selecting a predefined number of interesting tokens, the interesting tokens being
 - 23 the generated tokens having the greatest non-neutral probability values;
 - 24 (H) performing a Bayesian analysis on the selected interesting tokens to generate a
 - 25 spam probability; and
 - 26 (I) categorizing the email message as a function of the generated spam probability.

1 2. A method comprising the steps of:
2 receiving an email message comprising a text body having non-displaying
3 characters;
4 removing the non-displaying characters from the text body to generate a
5 displayable text body;
6 tokenizing the words in the displayable text body to generate tokens representative
7 of the displayable text body.

1 3. The method of claim 2, wherein the step of removing the non-displaying
2 characters comprises the step of:
3 removing non-displaying comment lines.

1 4. The method of claim 3, wherein the step of removing the non-displaying
2 characters comprises the step of:
3 removing non-displaying control characters.

1 5. The method of claim 4, wherein the step of removing the non-displaying
2 control characters comprises the step of:
3 removing characters associated with document format.

1 6. A method comprising the steps of:
2 receiving an email message comprising a text body, an SMTP email address, and
3 a domain name corresponding to the SMTP email address;
4 tokenizing the SMTP email address to generate a token representative of the
5 SMTP email address;
6 tokenizing the domain name to generate a token representative of the domain
7 name; and
8 determining a spam probability from the generated tokens.

1 7. The method of claim 6, further comprising the steps of:
2 removing non-displaying characters from the text body to generate a displayable
3 text body;
4 tokenizing the words in the displayable text body to generate tokens representative
5 of the displayable text body.

1 8. The method of claim 7, wherein the step of removing the non-displaying
2 characters comprises the step of:
3 removing non-displaying comment lines.

1 9. The method of claim 7, wherein the step of removing the non-displaying
2 characters comprises the step of:
3 removing non-displaying control characters.

1 10. The method of claim 9, wherein the step of removing the non-displaying
2 control characters comprises the step of:
3 removing characters associated with document format.

1 11. The method of claim 6, wherein the step of determining the spam
2 probability comprises the steps of:
3 assigning a spam probability value to the token representative of the SMTP email
4 address;
5 assigning a spam probability value to the token representative of the domain
6 name; and
7 generating a Bayesian probability value using the spam probability values
8 assigned to the tokens.

1 12. The method of claim 11, wherein the step of determining the spam
2 probability further comprises the step of:
3 comparing the generated Bayesian probability value with a predefined threshold
4 value.

1 13. The method of claim 12, wherein the step of determining the spam
2 probability further comprises the step of:
3 categorizing the email message as spam in response to the Bayesian probability
4 value being greater than the predefined threshold.

1 14. The method of claim 12, wherein the step of determining the spam
2 probability further comprises the step of:
3 categorizing the email message as non-spam in response to the Bayesian
4 probability value being not greater than the predefined threshold.

1 15. A method comprising the steps of:
2 receiving an email message comprising an attachment;
3 tokenizing the attachment to generate a token representative of the attachment;
4 and
5 determining a spam probability from the generated token.

1 16. The method of claim 15, wherein the step of receiving the email message
2 further comprises the step of:
3 receiving an email message including a text body.

1 17. The method of claim 16, further comprising the step of:
2 tokenizing the words in the text body to generate tokens representative of the
3 words in the text body.

1 18. The method of claim 17, wherein the step of tokenizing the words in the
2 text body comprises the steps of:
3 removing non-displaying characters from the text body to generate a displayable
4 text body;
5 tokenizing the words in the displayable text body to generate tokens representative
6 of the displayable text body.

1 19. The method of claim 17, wherein the step of determining the spam
2 probability comprises the steps of:
3 assigning a spam probability value to each of the tokens representative of the
4 words in the text body;
5 assigning a spam probability value to the token representative of the attachment;
6 and
7 generating a Bayesian probability value using the spam probability values
8 assigned to the tokens.

1 20. The method of claim 19, wherein the step of determining the spam
2 probability further comprises the step of:
3 comparing the generated Bayesian probability value with a predefined threshold
4 value.

1 21. The method of claim 20, wherein the step of determining the spam
2 probability further comprises the step of:
3 categorizing the email message as spam in response to the Bayesian probability
4 value being greater than the predefined threshold.

1 22. The method of claim 20, wherein the step of determining the spam
2 probability further comprises the step of:
3 categorizing the email message as non-spam in response to the Bayesian
4 probability value being not greater than the predefined threshold.

1 23. A system comprising:
2 email receive logic configured to receive an email message comprising an SMTP
3 email address and a domain name corresponding to the SMTP email address;
4 tokenize logic configured to tokenize the SMTP email address to generate a token
5 representative of the SMTP email address;
6 tokenize logic configured to tokenize the domain name to generate a token
7 representative of the domain name; and
8 analysis logic configured to determine a spam probability from the generated
9 tokens.

1 24. A system comprising:
2 means for receiving an email message comprising an SMTP email address and a
3 domain name corresponding to the SMTP email address;
4 means for tokenizing the SMTP email address to generate a token representative
5 of the SMTP email address;
6 means for tokenizing the domain name to generate a token representative of the
7 domain name; and
8 means for determining a spam probability from the generated tokens.

1 25. A computer-readable medium comprising:
2 computer-readable code adapted to instruct a programmable device to receive an
3 email message comprising an SMTP email address and a domain name corresponding to
4 the SMTP email address;
5 computer-readable code adapted to instruct a programmable device to tokenize the
6 SMTP email address to generate a token representative of the SMTP email address;
7 computer-readable code adapted to instruct a programmable device to tokenize the
8 domain name to generate a token representative of the domain name; and
9 computer-readable code adapted to instruct a programmable device to determine a
10 spam probability from the generated tokens.

1 26. The computer-readable medium of claim 25, further comprising:
2 computer-readable code adapted to instruct a programmable device to assign a
3 spam probability value to the token representative of the SMTP email address;
4 computer-readable code adapted to instruct a programmable device to assign a
5 spam probability value to the token representative of the domain name; and
6 computer-readable code adapted to instruct a programmable device to generate a
7 Bayesian probability value using the spam probability values assigned to the tokens.

1 27. The computer-readable medium of claim 26, further comprising:
2 computer-readable code adapted to instruct a programmable device to compare the
3 generated Bayesian probability value with a predefined threshold value.

1 28. The computer-readable medium of claim 27, further comprising:
2 computer-readable code adapted to instruct a programmable device to categorize
3 the email message as spam in response to the Bayesian probability value being greater
4 than the predefined threshold.

1 29. The computer-readable medium of claim 27, further comprising:
2 computer-readable code adapted to instruct a programmable device to categorize
3 the email message as non-spam in response to the Bayesian probability value being not
4 greater than the predefined threshold.

1 30. A system comprising:
2 email receive logic configured to receive an email message comprising an
3 attachment;
4 tokenize logic configured to tokenize the attachment to generate a token
5 representative of the attachment; and
6 analysis logic configured to determine a spam probability from the generated
7 token.

1 31. A system comprising:
2 means for receiving an email message comprising an attachment;
3 means for tokenizing the attachment to generate a token representative of the
4 attachment; and
5 means for determining a spam probability from the generated token.

1 32. A computer-readable medium comprising:
2 computer-readable code adapted to instruct a programmable device to receive an
3 email message comprising an attachment;
4 computer-readable code adapted to instruct a programmable device to tokenize the
5 attachment to generate a token representative of the attachment; and
6 computer-readable code adapted to instruct a programmable device to determine a
7 spam probability from the generated token.

1 33. The computer-readable medium of claim 32, further comprising:
2 computer-readable code adapted to instruct a programmable device to receive an
3 email message having a text body.

1 34. The computer-readable medium of claim 33, further comprising:
2 computer-readable code adapted to instruct a programmable device to tokenize the
3 words in the text body to generate tokens representative of the words in the text body.

1 35. The computer-readable medium of claim 34, further comprising:
2 computer-readable code adapted to instruct a programmable device to assign a
3 spam probability value to each of the tokens representative of the words in the text body;
4 computer-readable code adapted to instruct a programmable device to assign a
5 spam probability value to the token representative of the attachment; and
6 computer-readable code adapted to instruct a programmable device to generate a
7 Bayesian probability value using the spam probability values assigned to the tokens.

1 36. The computer-readable medium of claim 35, further comprising:
2 computer-readable code adapted to instruct a programmable device to compare the
3 generated Bayesian probability value with a predefined threshold value.

1 37. The computer-readable medium of claim 36, further comprising:
2 computer-readable code adapted to instruct a programmable device to categorize
3 the email message as spam in response to the Bayesian probability value being greater
4 than the predefined threshold.

1 38. The computer-readable medium of claim 36, further comprising:
2 computer-readable code adapted to instruct a programmable device to categorize
3 the email message as non-spam in response to the Bayesian probability value being not
4 greater than the predefined threshold.